Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Challenges in data science: a complex systems perspective

Anna Carbone[a,*], Meiko Jensen[b], Aki-Hiro Sato[c]

[a] Politecnico di Torino, Italy
[b] University of Southern Denmark, Odense, Denmark
[c] Kyoto University, Japan

ARTICLE INFO

Article history:
Received 25 April 2016
Accepted 29 April 2016
Available online 6 May 2016

PACS:
01.75.+m
05.10.-a
07.05.Kf
89.20.Hh

Keywords:
Big data
Disordered systems
Complex systems
Nonlinear systems

ABSTRACT

The ability to process and manage large data volumes has been proven to be not enough to tackle the current challenges presented by “Big Data”. Deep insight is required for understanding interactions among connected systems, space- and time- dependent heterogeneous data structures. Emergence of global properties from locally interacting data entities and clustering phenomena demand suitable approaches and methodologies recently developed in the foundational area of Data Science by taking a Complex Systems standpoint. Here, we deal with challenges that can be summarized by the question: “What can Complex Systems Science contribute to Big Data? ”. Such question can be reversed and brought to a superior level of abstraction by asking “What Knowledge can be drawn from Big Data?” These aspects constitute the main motivation behind this article to introduce a volume containing a collection of papers presenting interdisciplinary advances in the Big Data area by methodologies and approaches typical of the Complex Systems Science, Nonlinear Systems Science and Statistical Physics.

1. Introduction

Continuous developments in information and communications technology (ICT) have enabled the transition from an industrial society to an information society. The by-products of this transformation are huge amounts of data available to individuals, communities and institutions through the IT infrastructure itself [1-7]. The growing increase of digital data can be dated to 1990’s, when Internet had just begun to be used for commercial purposes. Since then, large amounts of digital data have been generated, stored and still grow in the cyber space. The yearly amount of digital data expected to be generated in 2020 is predicted as 40 Z Bytes according to the report published by the International Data Corporation [8].

The fundamental structure of the information society has been scrutinized at its early stage by pioneering new areas of data-intensive interdisciplinary approaches and methodologies [9–13]. New fields have thus emerged such as econophysics, sociophysics aimed at data-centric investigations of “stylized facts” in finance and social sciences [14–24].

Data of current interest (so-called “Big Data”) lack a predefined structure as they do not emerge from either a phenomenological model or an organized storing hierarchy. They are quite always inaccurate, out of context, inconsistent, imprecise, sparse thus needing even more data in order to be relevant in any investigation framework [25–28]. Lack of high-quality datasets has been recognized as the main reason of delayed development of many Artificial Intelligence (AI) breakthroughs [29]. In the first column of Table 1, AI advances are listed together with their related algorithms (second column), training databases (third column) and years of availability. An average delay of about 18 years can be observed between algorithm availability and breakthrough completion, whereas the average elapsed time with respect to key dataset availability is about three years, meaning that the lack of suitable datasets might have delayed the fulfillment of the breakthrough.

The suitability of key datasets and related technologies could become even more critical for the *Internet of Things* (IoT) deployment that requires the *Artificial Intelligence* to be upgraded to the level of *Ambient Intelligence* and adapted to the physical world. Imagine the multiplicity and heterogeneity of datasets needed to operate self-driving cars moving in cities or highways, autonomous robots assisting humans in everyday life and all those situations where machine prevails over human control. The IoT ecosystems, where humans and autonomous systems interact, require unconventional approaches to tackle challenges emerging from heterogeneous datasets and systems. Time series, spatial maps and images, ground weather stations, satellite-based data, each growing at

[*] Corresponding author.
E-mail address: anna.carbone@polito.it (A. Carbone).

http://dx.doi.org/10.1016/j.chaos.2016.04.020
0960-0779/© 2016 Elsevier Ltd. All rights reserved.

Table 1
Artificial Intelligence Breakthroughs (first column), Algorithms (second column) and Datasets (third column) [29].

AI advance	*Algorithm*	*Dataset*
Speech recognition (1994)	Hidden Markov Model (1984)	Spoken Wall Street journal articles (1991)
IBM Deep Blue chess player (1997)	Nega scout planning (1987)	Grandmaster chess games (1991)
Google Arabic and Chinese translation (2005)	Statistical machine translation (1988)	Google web news pages (2005)
IBM's Watson World Jeopardy champion (2011)	Mixture of experts (1991)	Wikipedia, Wiktionary, Wikiquote and Project Gutenberg (2010)
GoogleNet near-human performance object classification (2014)	Convolutional neural network (1989)	ImageNet datasets (2010)
Google Deep Mind human parity in playing Atari games (2015)	Deep Q-network-learning (1982)	Arcade learning environment (2013)

different spatial or temporal rate, should be integrated despite their heterogeneity.

2. What knowledge can be drawn from big data?

According to the scenario described in the Introduction, increasingly nested information impacting on decisions and services, intrinsic and extrinsic sustainability and resilience of connected multidimensional data systems are progressively becoming common issues for policy makers, economists and scientists. Traditional analytic may be inadequate to provide useful insights due to ever growing size and heterogeneity of collected data with decreased information gain. Computational methods need to be developed to quantify and characterize simultaneous and mutually interacting evolutions of autonomous scenarios over multiple:

- spatio-temporal scales (e.g. individual, local, urban, regional, global)
- dimensions (e.g. communication-financial-road-energy infrastructural multilayered networks)

2.1. The alphabet of big data

Initially Big Data issues were just limited to the effective handling of large amount of data. Hence, the only relevant feature to their description was *Volume*, which was referred to the utter amount of the dataset to be processed. The bigger data has been collected, the more representative the results can become, but also the more storage capacities and computational resources need to be provided for implementing the big data analytic. However, it was soon realized that *Volume* was not enough and a few more v's properties were introduced [30,31]:

Velocity refers to the speed at which data is produced. As new data items come in over time, a crucial parameter of big data systems design consists in anticipating the rate of items per second to be processed, stored and/or analyzed. Depending on the domain of application, the sheer velocity may largely impact the feasibility of big data analytic.

Variety refers to the heterogeneity arising from the integration of different datasets, following different schemes, structures and scales. Harmonizing these manifold data representations into a useful, linkable, processable homogeneous dataset is a complex challenge of its own.

Veracity refers to the uncertainty of data. Sensors may fail or become inaccurate, falsifying the dataset, and thereby impacting on the utility of the data in the analytics part. Data may get lost in transit, may become corrupted, deleted, altered, or just be falsely collected to begin with.

Volume, velocity, variety and veracity are known as the "*Four v's of Big Data*" (Fig. 1).

Since the pioneering work by Laney [30], the number of "Big Data v's" has steadily continued to grow. *Volatility* has been introduced as an estimate of the second-order moment of big data variety, in analogy with the meaning of volatility widely used in finance, that refers to the second-order moment of prices fluctuations and is roughly estimated as a windowed variance of return.

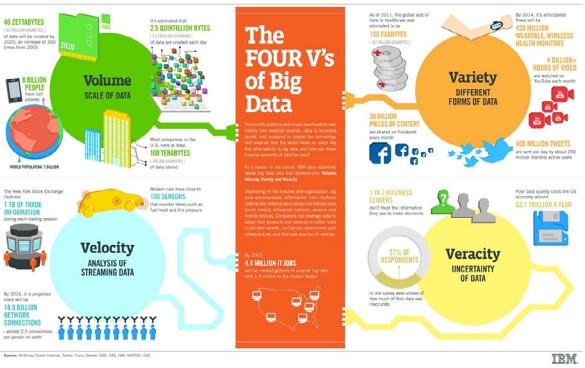
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
6 BILLION PEOPLE have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Fig. 1. The four v's of Big Data (courtesy of [31]).

Data on human interactions are highly volatile by nature, and do not even follow a consistent data scheme to be assumed for crafting the processing algorithms.

Value has been introduced to point to the relevance that Big Data might have for individuals, institutions and whole society with its impact on people wealth, health, security.

The list could become much longer: *vulnerability, viscosity, virality* and many other v's have been envisaged. The aim of enumerating the Big Data v's was to outline that a nested hierarchy of measurable interlinked properties is needed by the Big Data community at large. A Data Science is needed to deal with the challenge of extracting meaningful and valuable information (knowledge) from those properties.

2.2. Towards a science of big data

From the above introduction, it appears evident that the interest towards Big Data does not just rely in the ability to process and manage large data volumes. Rather, it is the knowledge potential unleashed by integrating several sources of data than ever before structured and unstructured data, new and old data, big and small data, environmental and behavioral data [2]. This is known as the "Variety Challenge" and is considered as the top data priority according to the fourth annual Big Data Executive Survey, conducted by NewVantage Partners [32]. In this section, we will try to discuss and clarify why a foundational data science is needed to gain a more profound awareness of the complex socioeconomic-technological-environmental systems of current interest. Terminology and concepts illustrated in the previous sections will be extended to ground them in more traditional scientific fields such as the Statistical Physics of Nonequilibrium and Nonlinear systems, Complexity and Social Sciences.

Fig. 2 shows a two dimensional graph with the *x* and *y*-axis corresponding respectively to "approach" and "methodology". For the sake of simplicity, bottom-up and top-down approaches have been assumed as the extremes of the x-axis, whereas human-made and cyber-enabled methodologies have been assumed as the extremes of the y-axis.